

EHSC Data Management and Sharing Plan*

The PI/PD is committed to sharing data produced by the University of Rochester (UR) Environmental Health Sciences Center (EHSC) members, and to disseminating results and information in a timely fashion and as openly as possible. We agree to manage and share materials and data that arise in pursuit of our goals and outline here seven elements of our plans to manage and share data, including metadata.

Element 1: Data Type

A. Data types: Data types of Center investigators can take almost any form. Several methods of data management and sharing may be used, including publication in peer-reviewed journals, documentation of research data repositories and archives, and dissemination of findings from specific research studies through other means. Data will be archived in the form(s) and type(s) of files appropriate to the methods that generated the data. For example, research protocols will generally be archived in a text format, such as MS Word or PDF, and research results will, to the extent possible and allowable, be shared in open formats, such that specialized software is not needed to access the data.

The types of data generated in the pursuit of EHSC member research projects and Center goals will vary depending upon the particular research. **Table 1** provides examples of the types of data that might be generated with support of this P30 Core Center grant. In some instances, use of data is straightforward, and no additional documentation or agreements are required. However, it may be necessary to utilize more sophisticated means to share some data, such as data from genome-wide transcription or chromatin accessibility analyses, data regarding levels of specific genes, proteins, histological and immunohistochemical measures, biochemical assays, etc., which use different formats for recording and transmitting data. In some cases, the data, or some aspects of data collected, may be confidential and contain protected health information. In such cases, data use agreements will need to be executed and approved prior to data sharing, and data shared in a de-identified format. In each case, the investigator will insure, to the extent that is feasible, that the research study data created in pursuit of the EHSC's goals or using Center supports and services are shared in commonly used formats. When this is not possible, the data will be shared as direct output of files from the specialized equipment that generated the data. Since a dataset by itself is likely to be of little value without the knowledge of the study details and the context under which the research operated, the datasets will be associated with necessary metadata, which are archived along with a related publication, research protocol or other documentation of the original research.

Individual faculty who are principal investigators (PIs) of research grants will be responsible for research data management in accordance with the guidelines and policies of agencies that fund this work (e.g., NIH, NSF). Based on current scope of EHSC member research, we estimate that data files will use up to 100 TB of storage. In addition to data, associated metadata, including coded indicators of each sample, will be archived. This P30 may in some instances, such as via the funding of Pilot Projects, support the creation of new information and data. These data will be shared at the time of peer reviewed publication and/or at the conclusion of this award, whichever comes first. These data will be handled following this data management and sharing plan, **Table 1** illustrates several examples of the types of data that may be generated and repositories that may be used. Elements 2-7 provide an overview of data management and sharing plans.

Also, there will be other information collected solely related to EHSC operations, including core usage, publications, member funding, collaborations, surveys and metadata to make this information easily interpretable and shareable. We estimate that this will use up <1GB of storage, will primarily be xls, txt and csv files, and will be archived in UR managed secure servers, REDcap and URBox. We do not anticipate that Center operations data will be published or made public; however, this information will be shared with Center leadership, Center members and the funding agency as part of annual progress reporting and continuous program evaluation.

B. Data preservation and backup: Maintaining records related to the Center will allow the PI/PD to document Center accomplishments and contributions. Additionally, data will be archived using both external hard drives and on locally maintained and regularly backed up servers (maintained by the University). All center operations data will be archived in an encrypted system, and accessible only by authorized personnel. In addition, the PI/PD and all EHSC members have access to UR Research, which is a web-based institutional repository system

based on a relational database, with collaborative authoring and versioning capabilities, and persistent identifiers. UR Research provides a password protected private workspace that allows UR users to upload and store up to 10 GB of files of any type. UR Research can handle file uploads in the 200-300 MB range. All files uploaded have a checksum calculated upon upload for file integrity verification. Key features include:

- Each submission to UR Research is assigned a DOI that enables easy search and retrieval.
- UR Research allows for the storage and retrieval of multiple versions of any given file. Registered users may share files with collaborators, within or outside of the UR, for co-authoring purposes. Collaborators must create an account in UR Research and always log in to access materials being shared with them. A publication area is also provided and can be used to disseminate material to larger groups. The publication area allows the content to be restricted to a given set of users when necessary. This progression from a work-in-progress file uploaded to the private workspace, to the sharing of the file with co-authors or peers for feedback, to the final publication of the work, parallels the data acquisition, analysis, and publication process.
- When an author publishes a file or files into the UR Research public repository, the record is automatically assigned a persistent identifier. This address is guaranteed to always return users to that original work, regardless of changes in the software or hardware being used to run the system. The system is run over a secure connection and uses a signed certificate that encrypts and protects stored information. The current certificate is supplied by Verisign and is encrypted at 128 AES.

Possible types of data	Data format	File sizes	Data repositories	Data sharing
Biochemical, molecular and cellular measurements (e.g., flow cytometry, polymerase chain reaction, enzymatic activity assays)	Txt csv, fcs, rdml, xlsx, imdx, d, raw	10MB-1 GB	Figshare Dryad LabArchives URResearch DMP Tool R/bioconductor	Metadata will be accessible and shared in common and open formats alongside datasets. Any computer code for computational procedures developed for data analysis will be made available.
Imaging microscopy (e.g., confocal microscopy, immunohistochemistry)	nd2, lif, czi, tiff, jpg, bmp, gif, pdf, xlsx, csv, ome-zarr	100MB-50 GB	Figshare Dryad LabArchives URResearch DMP Tool R/Bioconductor	Most manufacturers' of specialized imaging equipment primarily output proprietary files, but imaging tools such as ImageJ/Fiji are used to translate and view these files in open standard formats. Metadata will be accessible and shared in common and open formats alongside datasets. When used, unique computer code for data processing and analyses will be made available.
Analysis of organic or non-organic molecules or atoms in specimens (e.g., ICP-MS, LC/LC MS), which could include biological materials (e.g., cells or fluids from experimental studies) or environmental materials. (e.g., water, air sampling).	csv, xlsx, imdx, d, raw, wiff, mzML,	10-50 MB	Figshare Dryad LabArchives URResearch DMP Tool R/bioconductor	Metadata will be accessible in common and open formats alongside datasets. Any computer code developed for data processing or analysis will be made available.
"Omics" data (e.g., genomic, proteomic metabolomic analyses)	Fastq, fasta, SAM, BAM, mzXML, mzML	1GB-1TB	NCBI GEO MassIVE PRIDE R/bioconductor	Metadata will be accessible and shared in common and open formats alongside datasets. Proprietary data will be converted to open formats prior to depositing. Any computer code for computational procedures will be made available.
Survey data, and qualitative interview transcripts	.txt, .csv, pdf	<100 MB	UR BOX, RedCap	To protect privacy and confidentiality, shared data will be de-identified, following

				HIPAA and, when relevant, international Human Subject Protection laws.
--	--	--	--	--

- Data preservation is rigorously maintained. The UR Research system is backed up nightly. Also, 30 days of backups are maintained locally on disk; after 30 days the disks are written to tape, and tapes are rotated so there are always 12 months' worth kept off-site, in a fire and waterproof rated safe.
- The server holding the UR Research data is located in the UR River Campus Libraries' server room, which is temperature and humidity HVAC controlled. The server room is a secure, purpose-built room, accessible by keypad only by the libraries' systems administrators. The room is alarmed for access and for changes in temperature or humidity.

C. Metadata, other relevant information, and associated documentation: All Center members informed that they are required to follow current NIH guidelines and regulations regarding data management, data sharing and publishing, and are reminded of this upon receipt of a Center Pilot Project or other forms of Center support. In addition to data files, the methodology, software used, processed data, and summative results will be disclosed in publications. We also encourage EHSC members to publish using preprint servers to provide broad access to their research findings. Upon peer reviewed publication, data will be provided following the guidelines of the publisher and relevant funding agencies. This may include sharing raw or processed data via publicly accessible data repositories as well as in supplementary information associated directly with the research paper. If study data includes any Protected Health Information (PHI), then this will prevent open data sharing. As such, information will be shared in accordance with all privacy, ethical, and legal restrictions (e.g., Health Insurance Portability and Accountability Act or HIPAA), which in some instances prohibit public archiving. In this instance, the data will be shared on a restricted basis, and only after legal and other regulatory evaluations, processes and approvals have been completed. The specific repository will depend upon the type of data and other requirements (please see **Table 1** for examples). To facilitate the interpretation of the data deposited, metadata and protocols will be shared and associated with the applicable datasets. Data are often analyzed using commercially available software. Processed or summative data, the software used and, when appropriate any code used, will be provided. Unless proprietary data or processes are used, information and code will be made freely available, and data converted to an accessible file type.

Element 2: Related Tools, Software and/or Code

We will use the resources provided by the Edward G. Miner Library at the UR to deposit, store, and share raw and analyzed data generated. The [UR Research Repository \(URRR\)](#) provides a place for faculty, researchers, students, staff of the URochester to deposit their research outputs. The university provides access to software, including Figshare, LabArchives and DMPTools, and REDCap for data management and sharing. Optimized research protocols will be documented under *protocols.io* to store and share with other researchers if requested. Managing research study data that contains Protected Health Information (PHI) will be done with strict adherence to HIPAA and institutional regulations. Researchers must utilize secure, encrypted databases (e.g., REDCap), obtain IRB approval prior to any study, and the study will be required to include and use a system that controls access to the data by the internal research team, including collaborators at the University, and/or Data Use Agreements (DUAs) when sharing data with external collaborators. If any EHSC-supported research involves the use or generation of specialized tools, software, or codes then information about these will be shared and made available (or the data will be converted to an open use format prior to being deposited).

Element 3: Standards

Data types at the UR EHSC are investigator- and project-specific. The nature of the data and restrictions that apply will guide investigators on their choice of data preservation and sharing standards. The UR EHSC strongly encourages all investigators to share data and will ensure that appropriate methods are developed and used to achieve management and sharing of data in all appropriate situations. Data will be structured and described using standards that have been widely adopted in the field of study related to the specific research. For example, for flow cytometry, metadata generally include experimental parameters, cell type(s), machine identification, fluorescence minus one and singly stained samples, and a list of all antibodies, vendors, lots and fluorochromes. For analysis of chemical exposures and other agents (e.g., in biological specimens or environmental samples), NIST standard reference standards will be used. For methods for which positive and negative controls are available, relevant and feasible, they will be used and reported.

Element 4: Data Preservation, Access, and Associated Timelines

To ensure that the data generated in this project are preserved and broadly available to others, we will use established data repositories. Research data will be made available to others as soon as possible, and no later than at the time of an associated peer reviewed publication, or at the end of the award period (whichever comes first).

Repository where scientific data and metadata will be archived: Research data along with additional information that make the data interpretable and reusable (i.e., metadata) will be deposited in established repositories so that they are available to others. Deposited data receive a DOI which enables easy access and retrieval. Examples of metadata include, but are not limited to, treatment group identifiers, variables and variable descriptors, indication of methodology, and any applicable transformations. Examples of anticipated repositories are provided in **Table 1**. These include NCBI, UR Research, Dryad. Some types of data are deposited in generalist repositories such as FigShare or Dryad. If specialized tools, software, or code are used, then these will be identified and, when possible, data will be converted to open format prior to being deposited. The specific repository selected will depend upon the type of data and formats compatible with the repository. For example, if a Center member acquires sequencing data, such as via the pursuit of a funded Center Pilot Project, then these data and accompanying metadata will be submitted to the sequence read archive at NCBI for general use, and will be released (i.e., made publicly available) no later than the acceptance for publication of the main findings in accordance with NIH policies. Because all data sharing must adhere to all provisions of HIPAA, there may be some instances in which all study information may not be shareable due to containing PHI. Likewise, when used, data from third parties (e.g., electronic health records, census data, licensed data) may have restrictions on access, distribution and/or reuse, which include confidentiality protections. Data that are protected by law, confidential, proprietary or subject to non-disclosure agreements will not be shared, or only shared after receiving proper prior legal clearances and validated approvals.

A. How scientific data will be findable and identifiable: Electronic files are coded or indexed. Datasets that are available to share, as discussed above, will be identified and a DOI persistent identifier, and other information as allowed, such as project number, the PI's name, project title, and key words. In addition to a detailed methods section for publications associated with this work, Center members are requested to maintain prospectively prepared standard operation procedures, and share them openly. Also, in peer reviewed publications, detailed information for all methods used, including in supplemental materials, as well as indicating that data sharing must adhere to all provisions of HIPAA, which means that in some instances, data may not be shareable due to containing protected health information (PHI).

B. When and how long the scientific data will be made available: The research community and public at large will have access to research data at the time of publication. The data files will be stored in repositories supported and maintained by UR. Data will be retained by the investigator and UR for a minimum 5-year duration after publication.

Element 5: Access, Distribution, or Reuse Considerations

A. Whether access to scientific data will be controlled: Data access, sharing and reuse will abide by all provisions of HIPAA, the University, and by the agencies or entities the fund or regulate the specific work. Research data will be made available to others as soon as possible, and no later than at the time of an associated peer reviewed publication, or at the end of the award period (whichever comes first). The PI/PD will consider requests to share data prior to publication and/or prior to conclusion of this award on a case-by-case basis. If requests for data or other information related to Center member research project funded by sources other than (or in addition to) this P30 grant are made to the PI/PD of this P30, then the request will be forwarded to the person or group directly responsible for stewardship of that information.

B. Protections for privacy, rights, and confidentiality of human research participants: Sharing some data may require legal or other approvals, including but not limited to data use agreements, prior to sharing. In some instances, data may not be access, distributed or re-used due to the ability to identify individual

participants or infer the identity of participants. In all instances, research will adhere to all provisions of the HIPAA, which protects the privacy and security of patient and participant health information. The foundational principle of HIPAA is that the rights and privacy of all people who participate in research must be protected at all times. Data for sharing will be free from any identifiers that would link the results to any individual research participants. This may mean that some data, prior to being deposited, undergo a process of removing any PHI. Data will be stored deidentified in common and open formats. Information needed to make use of these data will be recorded in data dictionaries and codebooks that will be accessible to the research team and will subsequently be shared alongside final datasets. As such, any released data will be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual study participants.

Element 6: Oversight of Data Management and Sharing

The UR Office of Research and the Miner Library at the UR will be assisting and monitoring in data deposition and management during the duration of the project. The PI/PD of the EHSC will oversee the collection, management and sharing of data related to Center operations. This will include data on core and facility usage, publications, member funding, collaborations, internal annual and evaluative surveys of Center members. While we do not anticipate that these data will be published or made public, this information will be shared with Center members, University leadership (e.g., Vice Provost for Research, Dean to whom the Center director reports) and with the NIEHS as part of annual progress reporting. Research project data will be managed and archived by a research project data manager or laboratory technician within the relevant research project team. As such, for most of the data collected by Center members, the faculty PI of research grant that supports that study will oversee data collection, analysis, storage, and sharing. For studies that are funded internally, by this P30 grant (e.g., pilot projects), compliance will be monitored by the PI/PD of the P30 grant.

Element 7: Considerations related to animals and unique reagents

Although we do not anticipate that this P30 grant will directly generate any new strains of research animals, if this were to occur, they will be deposited in appropriate NIH-supported repository, if one exists. For example, for mice, they would be deposited in a mouse repository for the cryopreservation of embryos or sperm that would be available for distribution to biomedical researchers. If novel strains of mice are generated, then we will use standard nomenclature and receive approval from the Mouse Genome Informatics (MGI) nomenclature committee <http://www.informatics.jax.org/>. Following characterization and peer-reviewed publication of any novel reagents or research animals generated, they will be openly distributed to investigators at academic institutions who will use them for non-commercial research. Requests for novel animal strains or reagents created by others, including Center members, will be directed to the initial creator. Following publication of findings and appropriately addressing technology transfer requirements, resources will be shared, as available, to qualified academic investigators for non-commercial research. We have demonstrated our commitment to sharing by providing research animals (e.g., mice, frogs, fruit flies) and other novel reagents, after proper execution of appropriate Material Transfer Agreements, with numerous investigators over the past 50 years. If we assume responsibility for distributing a newly generated model organism, then we will fill requests in a timely fashion. In addition, we will provide relevant protocols and published genetic and phenotypic data upon request. Material transfers will be made with no more restrictive terms than in the Simple Letter Agreement (SLA) or the Uniform Biological Materials Transfer Agreement (UBMTA) and without reach through requirements. Should any intellectual property arise which requires a patent, we will ensure that the technology (materials and data) remains widely available to the research community in accordance with the NIH Principles and Guidelines document.

*Please refer to current NIH guidelines regarding Data Management and Sharing, as they may have changed since the preparation of this document.

<https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms>
<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-046.html>
<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>